UNIT 8 DIGITISATION CONCEPT AND NEED

Structure

8.0 OBJECTIVES

After going through this Unit, you will be able to:

- understand the basic concept, and need for digitisation;
- know how to select the materials for digitisation;
- explain the various steps involved in the process of digitisation; and
- enumerate the input and output options for digitisation.

8.1 INTRODUCTION

All recorded information in a traditional library is analogue in nature. The analogue information can include printed books, periodical articles, manuscripts, cards, photographs, vinyl disks, video and audio tapes. However, when analogue information is fed into a computer, it is broken down into 0s and 1s changing its characteristics from analogue to digital. These bits of data can be re-combined for manipulation and compressed for storage. Voluminous encyclopaedias that take-up yards of shelf-space in analogue form can fit into a small space on a computer drive or stored on to a CD ROM disc, which can be searched, retrieved, manipulated and sent over the network. One of the most important traits of digital information is that it is not fixed in the way that

texts printed on a paper are. Digital texts are neither final nor finite, and are not fixed either in essence or in form except when it is printed out as a hard copy.

Flexibility is one of the chief assets of digital information. An endless number of identical copies can be created from a digital file, because a digital file does not decay by copying. Moreover, digital information can be made accessible from remote location simultaneously by a large number of users. In this Unit, let us discuss about the definitions, basic concept, need, steps in the process of digitisation, etc.

8.2 DIGITISATION: DEFINITIONS

The word "digital" describes any system based on discontinuous data or events. Computers are digital machines because at their most basic level they can distinguish between just two values, 0 and 1, or off and on. All data that a computer processes must be encoded digitally as a series of zeroes and ones.

The opposite of digital is analogue. A typical analogue device is a clock in which the hands move continuously around the face. Such a clock is capable of indicating every possible time of the day. In contrast, a digital clock is capable .of representing only a finite number of times (every tenth of a second, for example).

As mentioned before, a printed book is an analogue form of information. The contents of a book need to be digitised to convert it into digital form. Digitisation is the process of converting the content of physical media (e.g., periodical articles, books, manuscripts, cards, photographs, vinyl disks, etc.) to digital formats.

Digitisation refers to the process of translating a piece of information such as a book, journal articles, sound recordings, pictures, audio tapes or video recordings, etc. into bits. Bits are the fundamental units of information in a computer system. Converting information into these binary digits (Bits) is called digitisation, which can be achieved through a variety of existing technologies. A digital image, in turn, is composed of a set of pixels (picture elements), arranged according to a pre-defined ratio of columns and rows. An image file can be managed as regular computer file and can be retrieved, printed and modified using appropriate software. Further, textual images can be OCRed so as to make its contents searchable.

Self Check Exercise

1) Define 'Digitisation'.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

8.3 NEED FOR DIGITISATION

Digitising a document in print or other physical media (e.g., sound recordings) makes the document more useful as well as more accessible. It is possible for a user to conduct a full-text search on a document that is digitised and OCRed. It is possible to create hyperlinks to lead a reader to related items within the text itself as well as to external resources. Ultimately, digitisation does not mean replacing traditional library collections and services; rather, it serves to enhance them.

A document can be converted into digital format depending on the objective of digitisation, end user, availability of finances, etc. While the objectives of digitisation initiatives differ from organisation to organisation, the primary objective is to improve access. Other objectives include cost savings, preservation, keeping pace with technology, and information sharing. The most significant challenges in planning and execution of a digitisation project relate to technical limitations, budgetary constraints, copyright considerations, lack of policy guidelines and, lastly, the selection of materials for digitisation.

While new and emerging technologies allow digital information to be presented in innovative ways, the majority of potential users are unlikely to have access to sophisticated hardware and software. Sharing of information among various institutions is often restricted by the use of incompatible software.

One of the main benefits of digitisation is to preserve rare and fragile objects with enhancing their access to multiple number of users simultaneously. Very often, when an object is rare and precious, access is only allowed for certain category of people. Going digital could allow more users to enjoy the benefit of access. Although, digitisation offers great advantages for access, allowing users to find, retrieve, study and manipulate material, it cannot be considered as a good alternate for preservation because of ever changing formats, protocols and software used for creating digital objects.

There are several reasons for libraries to go for digitisation and there are as many ways to create the digitised images, depending on the needs and uses. The prime reason for the digitisation is the need of the user for convenient access to high quality information. Other important considerations are:

Quality Preservation

The digital information has potential for qualitative preservation of information. The preservation-quality images can be scanned at high resolution and bit depth for best possible quality. The quality remains the same inspite of multiple usage by several users. However, caution need to be exercised while choosing digitized information as preservation media.

Multiple Referencing

Digital information can be used simultaneously by several users at a time.

Wide Area Usage

Digital information can be made accessible to distant users through the computer networks over the Internet.

Archival Storage

Digitisation is used for restoration of rare material. The rare books, images or archival materials are kept in digitised format as a common practice.

Security Measure

Valuable documents and records are scanned and kept in digital format for safety.

Self Check Exercise

2) Discuss the need for digitisation.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

8.4 SELECTION OF MATERIAL FOR DIGITISATION

To begin the process of digitisation, first of all, we need to select documents for digitisation. The process of selection of material for digitisation involves identification, selection and prioritisation of documents that are to be digitized. If an organisation generates contents, strategies may be adopted to capture data that is "born digital". If documents are available in digital form, it can be easily converted into other formats. If the selected material is from the external sources, IPR issues need to be resolved. It is important to obtain permission from the publishers and data suppliers for digitisation, if material being digitized is not available in public domain. Moreover, decision may be taken whether to OCR the digitized images. Documents selected for digitisation may already be available in digital format. It is always more economical to buy e-media, if available, than their conversion. Moreover, over-sized material, deteriorating collections, bound volumes of journals, manuscripts etc. would require highly specialised equipment and highly skilled manpower.

The documents to be digitized may include text, line art, photographs, colour images, etc. The selection of document, need to be reviewed very carefully considering all the factors of utility, quality, security and cost. Rare and highly required documents and images are given first priority in selection without considering the quality. Factors that may be considered before selecting different media for digitisation include:

Audio

The sound quality has to be checked and required corrections made together by the subject expert and computer sound editor.

Video

The video clippings are normally edited on Beta max tapes which can be used for transferring it on digital format. While editing colour tone and resolution is checked and corrected.

Photographs

The selection of photographs is very crucial process. High resolution is required for photographic images and slides. Especially the quality, future need and the copyright aspects have to be checked.

Documents

Documents which are much in demand, too fragile to handle, and rare in availability are reviewed and selected for the process. If the correction of literary value demands much input, then documents are considered for publication rather than digitisation.

Self Check Exercise

3) What are the factors to be considered before selecting different media for digitisation?

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

.....

8.5 STEPS IN THE PROCESS OF DIGITISATION

The following four steps are involved in the process of digitisation. Software, variably called document image processing (DIP), Electronic Filing System (EFS) and Document Management Systems (DMS) provides all or most of these functions:

8.5.1 Scanning

Electronic scanners are used for acquisition of an electronic image into a computer through its original that may be a photograph, text, manuscript, etc. An image is "read" or scanned at a predefined resolution and dynamic range. The resulting file, called "bit map page image" is formatted (image formats

describes elsewhere) and tagged for storage and subsequent retrieval by the software package used for scanning. Acquisition of image through fax card, electronic camera or other imaging devices is also feasible. However, image scanners are most important and most commonly used component of an imaging system for transfer of normal paper-based documents.

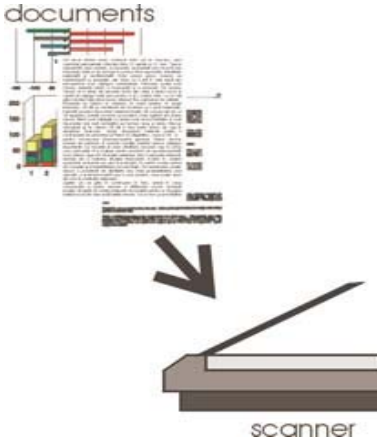


Fig. 8.l: Scanning a document using Flatbed Scanner

The following are the steps involved in the process of scanning using a flatbed scanner

- Step 1. Place picture on the scanner's glass

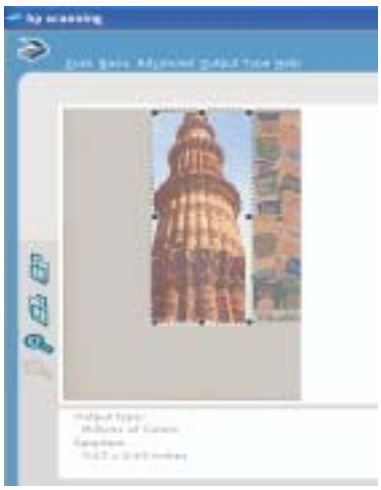
- Step 2. Start scanner software

- Step 3. Select the area to be selected for storage

- Step 4. Choose the image type

- Step 5. Sharpen the image

- Step 6. Set the image size



- Step 7. Save the scanned image using a desirable format (GIF or JPEG)

8.5.2 Indexing

If converting a document into an image or text file is considered as the first step in the process of imaging, indexing these files comprises the second step. The process of indexing scanned image involves linking of database of scanned images to a text database. Scanned images are just like a set of pictures that need to be related to a text database describing them and their contents. An imaging system typically stores a large amount of unstructured data in a two file system for storing and retrieving scanned images. The first is traditional file that has a text description of the image (keywords or descriptors) along with a key to a second file. The second file contains the document location. The user selects a record from the first file using a search algorithm. Once the user selects a record, the application keys into the location index, finds the document and displays it.

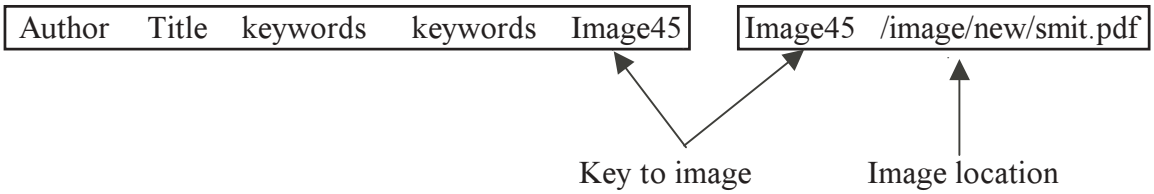


Fig. 8.2: Two File System in an Image Retrieval System

Most of the document imaging software packages, through their menu driven or command driven interface facilitate elaborate indexing of documents. While some document management system facilitates selection of indexing terms from the image file, others allow only manual keying in of indexing terms. Further, many DMS packages provide OCRed capabilities for transforming the images into standard ASCII files. The OCRed text then serves as a database for full-text search of the stored images.

8.5.3 Storing

The most tenacious problem of a document image relates to its file size and, therefore, to its storage. Every part of an electronic page image is saved regardless of present or absence of ink. The file size varies directly with scanning resolution, the size of the area being digitized and the style of graphic file

format used to save the image. The scanned images, therefore, need to be transferred from the hard disc of scanning workstation to an external large capacity storage devices such as an optical disc, CD-ROM / DVD-ROM disc, snap servers, etc. While the smaller document imaging system may use offline media, which need to be reloaded when required, or fixed hard disc drives allocated for image storage, larger document management systems use auto-changers such as optical jukeboxes and tape library systems. The storage required by the scanned image varies and depends upon factors such as scanning resolution, page size, compression ratio and page content. Further, the image storage device may be either remote or local to the retrieval workstation depending upon the imaging systems and document management systems used.

8.5.4 Retrieving

Once scanned images and OCRed text documents have been saved as a file, a database is needed for selective retrieval of data contained in one or more fields within each record in the database. Typically, a document imaging system uses at least two files to store and retrieve documents. The first is traditional file that has a text description of the image along with a key to the second file. The second file contains the document location. The user selects a record from the first-file using a search algorithm. Once the user selects a record, the application keys into the location index, finds the document and displays it. Most of the document management system provides elaborate search possibilities including use of Boolean and proximity operators (AND, OR, NOT) and wild cards. Users are also allowed to refine their search strategy. Once the required images have been identified their associated document image can quickly be retrieved from the image storage device for display or printed output.

Self Check Exercise

4) Write the steps involved in the process of scanning using a flatbed scanner.

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

8.6 DIGITISATION: INPUT AND OUTPUT OPTIONS

A document can be converted into digital format depending on the objective of digitisation, end user, availability of finance, etc. There are four basic approaches that can be adapted to convert from print to digital:

- Scanned as Image Only
- OCR and Retaining Page Layout
- Retaining Page Layout using Acrobat Capture; and
- Re-keying the Data

8.6.1 Scanned as Image Only

Image only is the lowest cost option in which each page is an exact replica of the original source document. Several digital library projects are concerned with providing digital access to materials that already exists with traditional libraries in printed media. Scanned page images are practically the only reasonable solution for institutions such as libraries and information centres for converting existing paper collection (legacy documents) without having access to the original data in computer processible formats convertible into HTML / SGML or in any other structured or unstructured text. Scanned page images are natural choice for large scale conversions for major digital library initiatives. Printed text, pictures and figures are transformed into computer-accessible forms using a digital scanner or a digital camera in a process called document imaging or scanning. The digitally scanned images are stored in a file as a bit-mapped page image, irrespective of the fact that a scanned page contains a photograph, a line drawing or text. A bit-mapped page image is a type of computer graphic, literally an electronic picture of the page which can very easily be equated to a facsimile image of the page and as such they can be read by humans, but not by the computers, understably "text" in a page image is not searchable on a computer using the present-day technology. An image-based implementation requires a large space for data storage and transmission.

Capturing page image format is comparatively easy and inexpensive, it is a faithful reproduction of its original maintaining page integrity and originality. The scanned textual images, however, are not searchable unless it is OCRed, which in itself, is highly error prone process especially when it involves scientific texts. Options and technology for converting print to digital are given separately.

Since OCR is not carried out, the document may not be searchable. Most of the scanning softwares generate TIFF format by default, which, can be converted into PDF using a number of software tools. Scan to TIFF / PDF format is recommended only when the requirement of project is to make documents portable and accessible from any computing platform. The image can be browsed through a table of contents file composed in HTML that provides link to scanned image objects.

8.6.2 OCR and Retaining Page Layout

The latest versions of both Xerox's TextBridge and Caere's Omnipage incorporate technology that allow the option of maintaining text and graphics in their original layout as well as plain ASCII and word-processing formats. Output can also include HTML with attributes like bold, underline, and italic retained.

Retaining Layout after OCR

A scanned document is nothing more than a picture of a printed page. It cannot be edited or manipulated or managed based on their contents. In other words, scanned documents have to be referred to by their labels rather than characters in the documents. OCR (Optical Character Recognition) programs are software tools used to transform scanned textual page images into word processing file. OCR or text recognition is the process of electronically identifying text in a bit-mapped page image or set of images and generate a file containing text in ASCII code or in a specified word processing format leaving the image intact in the process.

8.6.3 Retaining Page Layout using Acrobat Capture

The Acrobat Capture 2.0 provides several options for retaining not only the page layout but also the fonts, and to fit text into the exact space occupied in the original, so that the scanned and OCRed copy never over or under-shoots the page. Accordingly, it treats unrecognisable text as images that are pasted in its place. Such images are perfectly readable by anyone looking at the PDF file, but which will be absent from the editable and searchable text file. In contrast, ordinary OCR programs treat unrecognised text as tildes or some other special character in the ASCII output. Acrobat Capture can be used to scan pages as images, image +text and as normal PDF, all the three options retain page layout.

Image Only: Image only option has already been described in option 1.

Image + Text: In image+text solutions, a OCRed text is generated for each image where each page is an exact replica of the original and left untouched, however, the OCRed text sits behind the image and is used for searching. The OCRed text is generally not corrected for errors since it is used only for searching. The cost involved is much less than PDF Normal. However, the entire page is a bitmap and neither fonts nor line drawings are vectorised, so the file size of Image + Text PDFs is considerably larger than the corresponding PDF Normal files and pages will not display as quickly or cleanly on screen.

PDF Normal: PDF normal gives the clearest on-screen display, is searchable, and yet with significantly smaller file size than Image+Text. The result is not, however, an exact replica of the scanned page. While all graphics and formatting are preserved, substitute fonts may be used where direct matches are not possible. It is a good choice when files need to be posted to the web or otherwise delivered online. If, during the Capture and OCR process, a word cannot be recognised to the specified confidence level, Capture, by default, substitutes a small portion of the original bitmap image. Capture' "best guess" of the suspect word lies behind the bitmap so that searching and indexing are still possible. However, one cannot guarantee that these bitmapped words are correctly guessed. In addition, the bitmap is somewhat obtrusive, detracting from the 'look' of the page. Further, Capture provides option to correct suspected errors left as bit-mapped image or leave them untouched.

8.6.4 Re-keying the Date

A classic solution of this kind would comprise keying-in the data and its verification.

This involves a complete keying of the text, followed by a full rekeying by a different operator, the two keying-in operation might take place simultaneously. The two keyed files are compared and any errors or inconsistencies are corrected. This would guarantee at least 99.9% accuracy, but to reach 99.955% accuracy level, it would normally require full proof-reading of the keyed files, plus table lookups and dictionary spell checks.

Fig. 8.3: Rekeying-in as an Option for Digitisation

Self Check Exercise

5) What are the four basic approaches that can be adapted to translate from print to digital?

Note: i) Write your answer in the space given below.

ii) Check your answer with the answers given at the end of this Unit.

.....

.....

.....

.....

.....

.....

8.7 SUMMARY

Digitisation is the process of converting the content of physical media (e.g., periodical articles, books, manuscripts, cards, photographs, vinyl disks, etc.) into digital format. In most library applications, digitisation normally results in a documents that are accessible from the web site of a library, and thus on the Internet. Optical scanners and digital cameras are used to digitise images by translating them into bit maps. It is also possible to digitise sound, video, graphics, animations, etc.

Digitisation is the first step in the process of building digital libraries. Digitisation is also used for achieving preservation and archiving although it is not considered as good option for preservation and archiving. It is highly labour-intensive and cost-intensive process that involves several complexities including copyright and IPR issues. However, digital objects offer numerous

benefits in terms of accessibility and search. The documents to be digitised may include text, line art, photographs, colour images, etc. The selection of document, need to be reviewed very carefully considering all the factors of utility, quality, security and cost. Rare and much in demand documents and images are selected as first priority without considering the quality.

The process of digitisation invloves four steps namely, scanning, indexing, storage and retrieval. A scanned document is nothing more than a picture of a printed page. It cannot be edited or manipulated or managed based on their contents. In other words, scanned documents have to be referred to by their labels rather than characters in the documents. OCR (Optical Character Recognition) programs are software tools used to transform scanned textual page images into word processing file. OCR or text recognition is the process of electronically identifying text in a bit-mapped page image or set of images and generates a file containing that text in ASCII code or in a specified word processing format leaving the image intact in the process.

8.8 ANSWERS TO SELF CHECK EXERCISES

1) The word "digital" describes any system based on discontinuous data or events. Computers are digital machines because at their most basic level they can distinguish between just two values, 0 and 1, or off and on. All data that a computer processes must be encoded digitally as a series of zeroes and ones.

   The opposite of digital is analogue. A typical analogue device is a clock in which the hands move continuously around the face. Such a clock is capable of indicating every possible time of the day. In contrast, a digital clock is capable of representing only a finite number of times (every tenth of a second, for example).

   As mentioned before, a printed book is an analogue form of information. The contents of a book need to be digitised to convert it into digital form. Digitisation is the process of converting the content of physical media (e.g., periodical articles, books, manuscripts, cards, photographs, vinyl disks, etc.) to digital formats.

2) There are several reasons for libraries to go for digitisation and there are as many ways to create the digitised images, depending on the needs and uses. The prime reason for the digitisation is the need of the user for convenient access to high quality information. Other important considerations are:

   Quality Preservation

   The digital information has potential for qualitative preservation of information. The preservation-quality images can be scanned at high resolution and bit depth for best possible quality. The quality remains the same inspite of multiple usage by several users. However, caution need to be exercised while choosing digitised information as preservation media.

   Multiple Referencing

   Digital information can be used simultaneously by several users at a time.

Wide Area Usage

Digital information can be made accessible to distant users through the computer networks over the Internet.

Archival Storage

Digitisation is used for restoration of rare material. The rare books, images or archival materials are kept in digitised format as a common practice.

Security Measure

Valuable documents and records are scanned and kept in digital format for safety.

3) Factors that may be considered before selecting different media for digitisation include:

Audio

The sound quality has to be checked and required corrections made together by the subject expert and computer sound editor.

Video

The video clippings are normally edited on Beta max tapes which can be used for transferring it on digital format. While editing colour tone and resolution is checked and corrected.

Photographs

The selection of photographs is very crucial process. High resolution is required for photographic images and slides. Especially the quality, future need and the copyright aspects have to be checked.

Documents

Documents which are much in demand, too fragile to handle, and rare in availability are reviewed and selected for the process. If the correction of literary value demands much input, then documents are considered for publication rather than digitisation.

4) The following are the steps involved in the process of scanning using a flatbed scanner

- Step 1. Place picture on the scanner's glass
- Step 2. Start scanner software
- Step 3. Select the area to be selected for storage
- Step 4. Choose the image type
- Step 5. Sharpen the image
- Step 6. Set the image size
- Step 7. Save the scanned image using a desirable format (GIF or JPEG)

5) There are four basic approaches that can be adapted to translate from print to digital:

- Scanned as Image Only
- OCR and Retaining Page Layout
- Retaining Page Layout using Acrobat Capture; and
- Re-keying the Data

8.9 KEYWORDS

Archiving : The act of preparing paper or electronic documents for long-term storage so the contents are preserved, and can be used, if needed, at a later date.

Digital Collection : A collection of books, papers, photographs, etc. that are digitised and made available in electronic format.

Digitisation : The process of converting an image, text, or signal into digital code, usually by scanning or rekeying.

GIF (Graphics Interchange Format) : GIF files should only be used for online and/or viewing purposes limited to 256 colours.

HTML(Hyper Text Markup Language : HTML tags are used to build web pages; HTML defines the page layout, fonts and graphic elements as well as the hypertext links to pages on the website, or to other websites; HTML can be created in text-based editors, WYSIWIG programs, or by hand coding the tags in a word processor.

OCR : The process of converting the text from the archival image into that in the word processor.

SGML (Standard Generalized Markup Language) : The original markup language which uses tags to define content information.

XML (Extensible Markup Language) : It is used for defining data elements in documents and on web pages; it uses a similar tag structure as HTML; however, whereas HTML defines how elements are displayed, SML defines what those elements contain.

8.10 REFERENCES AND FURTHER READING

Arms, William Y. (2000). *Digital Libraries.* Cambridge, MA: The MIT Press.

Arms, W. Y. Key (1995). Concepts in the Architecture of the Digital Library. "*D-Lib Magazine*.

Kenney, Anne R. and Stephen, Chapman. (1996). *Digital Imaging for Libraries and Archives*. Ithaca: Dept. of Preservation and Conservation, Cornell University Library.

Kessler, Jack. (1996). *Internet Digital Libraries: The International Dimension.* Boston: Artech House Publishers.

Townsend, Sean (et. al.). *Digitising History*. (http://hds.essex.ac.uk!g2gp/digitising_history/index.html)